Gradient-Based Optimization

Lecture Notes by Dr. Suneesh Jacob for EIE417

1 Analytical Derivatives and Numerical Derivatives

1.1 Analytical Derivatives

For a given function $f(x)$, if x changes by $x+h$ then $f(x)$ would change by $f(x+h)$. If this change in x is infinitesimal (i.e., if h tends to zero), then the analytical derivative is given by

$$\frac{d}{dx}\left(f(x)\right) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{(x+h)-(h)} = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

If $f(x,y)$ is a function of two variables x and y, then the partial derivative of the function with respect to x would be

$$\frac{\partial}{\partial x}\left(f(x,y)\right) = \lim_{h\to 0} \frac{f(x+h,y)-f(x,y)}{h}$$

1.2 Numerical Derivatives

It is not very easy to compute analytical derivatives using computers (for that matter, it is not always easy to compute them manually either!), and hence we go for numerical derivatives.

For numerical derivative, instead of h being infinitesimally small, we take h to be finitely small, i.e., a very small number, such as $h = 0.001$, for example. This way, the numerical derivative can be computed as

$$\frac{d}{dx}\left(f(x)\right) \approx \frac{f(x+h)-f(x)}{h}$$

which does not require us to calculate the limit. But the derivative should be for infinitesimally small changes, not for finitely small changes. So, the result would not be exactly equal to the actual derivative, but would be numerically close to the actual derivative at a given point. This is typically referred to as Forward Difference Method.

Likewise, numerical partial derivative of a function $f(x,y)$ with respect to x would be

$$\frac{\partial}{\partial x}\left(f(x,y)\right) \approx \frac{f(x+h,y)-f(x,y)}{h}$$

1.3 Numerical Second-Order Partial Derivatives

The numerical second-order partial derivative of the function $f(x,y)$ with respect to x (twice) is given by

$$\frac{\partial^2}{\partial x^2}\left(f(x,y)\right) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\left(f(x,y)\right)\right) \approx \frac{\frac{\partial f}{\partial x}|_{(x+h,y)} - \frac{\partial f}{\partial x}|_{(x,y)}}{h}$$

$$\approx \frac{\frac{f(x+h+h,y)-f(x+h,y)}{h} - \frac{f(x+h,y)-f(x,y)}{h}}{h}$$

$$\approx \frac{f(x+2h,y) - 2f(x+h,y) + f(x,y)}{h^2}$$

The numerical second-order mixed partial derivative with respect to x and y is given by

$$\frac{\partial^2}{\partial x \partial y}\left(f(x,y)\right) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\left(f(x,y)\right)\right) \approx \frac{\frac{\partial f}{\partial y}|_{(x+h,y)} - \frac{\partial f}{\partial y}|_{(x,y)}}{h}$$

$$\approx \frac{\frac{f(x+h,y+h)-f(x+h,y)}{h} - \frac{f(x,y+h)-f(x,y)}{h}}{h}$$

$$\approx \frac{f(x+h,y+h) - f(x+h,y) - f(x,y+h) + f(x,y)}{h^2}$$

2 Gradient

For a function $f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$, the gradient $\nabla_{\boldsymbol{x}} f$ is defined as shown below

$$\nabla_{\boldsymbol{x}} f = \begin{Bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{Bmatrix} \tag{1}$$

For a function of two variables, i.e., $f(x_1, x_2)$, the gradient is given by

$$\nabla_{\boldsymbol{x}} f = \begin{Bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{Bmatrix} \tag{2}$$

2.1 Computation of the gradient of a given function at a given point

Problem statement: Suppose $f(x_1, x_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$. Find the gradient of the function at $(x_1, x_2) = (2, 3)$.

Solution:

$$\nabla_{\boldsymbol{x}} f = \begin{Bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{Bmatrix} = \begin{Bmatrix} 2\left(x_1^2 + x_2 - 11\right) \cdot 2x_1 + 2\left(x_1 + x_2^2 - 7\right) \\ 2\left(x_1^2 + x_2 - 11\right) + 2\left(x_1 + x_2^2 - 7\right) \cdot 2x_2 \end{Bmatrix}$$

At $x = (2, 3)$, the gradient would be

$$\begin{Bmatrix} 2\left((2)^2 + 3 - 11\right) \cdot 2 \cdot 2 + 2\left(2 + (3)^2 - 7\right) \\ 2\left((2)^2 + 3 - 11\right) + 2\left(2 + (3)^2 - 7\right) \cdot 2(3) \end{Bmatrix} = \begin{Bmatrix} -24 \\ 40 \end{Bmatrix}$$

2.2 Numerical computation of the gradient of a given function at a given point

Problem statement: Suppose $f(x_1, x_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$. Find the numerical gradient of the function at $(x_1, x_2) = (2, 3)$ by using Forward Difference Method with a step-size of $h = 0.001$.

Solution:

At $x = (2, 3)$ and with $h = 0.001$, the numerical gradient with Forward Difference Method would be

$$\nabla_{\boldsymbol{x}} f = \begin{Bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{Bmatrix} = \begin{Bmatrix} \frac{f(x_1+h, x_2) - f(x_1, x_2)}{h} \\ \frac{f(x_1, x_2+h) - f(x_1, x_2)}{h} \end{Bmatrix} = \begin{Bmatrix} \frac{f(2.001, 3) - f(2,3)}{0.001} \\ \frac{f(2, 3.001) - f(2,3)}{0.001} \end{Bmatrix} = \begin{Bmatrix} -23.991 \\ 40.045 \end{Bmatrix}$$

3 Hessian

For a function $f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$, the Hessian $\nabla_{\boldsymbol{xx}} f$ is defined as shown below

$$\nabla_{\boldsymbol{xx}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 x_n} & \frac{\partial^2 f}{\partial x_2 x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{3}$$

For a function of two variables, i.e., $f(x_1, x_2)$, the Hessian is given by

$$\nabla_{\boldsymbol{xx}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} \tag{4}$$

3.1 Computation of Hessian of a given function at a given point

Problem statement: Suppose $f(x_1, x_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$. Find the gradient of the function at $(x_1, x_2) = (2, 3)$.

Solution:

$$\nabla_{\boldsymbol{xx}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 + 4x_2 - 42 & 4x_1 + 4x_2 \\ 4x_1 + 4x_2 & 4x_1 + 12x_2^2 - 26 \end{bmatrix}$$

At $x = (2, 3)$, the Hessian would be

$$\begin{bmatrix} 12(2)^2 + 4(3) - 42 & 4(2) + 4(3) \\ 4(2) + 4(3) & 4(2) + 12(2)^2 - 26 \end{bmatrix} = \begin{bmatrix} 18 & 20 \\ 20 & 90 \end{bmatrix}$$

3.2 Numerical computation of the Hessian of a given function at a given point

Problem statement: Suppose $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$. Find the numerical Hessian of the function at $(x_1, x_2) = (2, 3)$ by using Forward Difference Method with a step-size of $h = 0.001$.

Solution:

At $x = (2, 3)$ and with $h = 0.001$, the numerical Hessian with Forward Difference Method would be

$$\nabla_{\boldsymbol{xx}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} \\ \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} \end{bmatrix} = \begin{bmatrix} \frac{f(x_1+2h,x_2)-2f(x_1+h,x_2)+f(x_1,x_2)}{h^2} & \frac{f(x_1+h,x_2+h)-f(x_1+h,x_2)-f(x_1,x_2+h)+f(x_1,x_2)}{h^2} \\ \frac{f(x_1+h,x_2+h)-f(x_1+h,x_2)-f(x_1,x_2+h)+f(x_1,x_2)}{h^2} & \frac{f(x_1,x_2+2h)-2f(x_1,x_2+h)+f(x_1,x_2)}{h^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{f(2.002,3)-2f(2.001,3)+f(2,3)}{0.001^2} & \frac{f(2.001,3.001)-f(2.001,3)-f(2,3.001)+f(2,3)}{0.001^2} \\ \frac{f(2.001,3.001)-f(2.001,3)-f(2,3.001)+f(2,3)}{0.001^2} & \frac{f(2,3.002)-2f(2,3.001)+f(2,3)}{0.001^2} \end{bmatrix} = \begin{bmatrix} 18.048 & 20.004 \\ 20.004 & 90.072 \end{bmatrix}$$

4 Gradient-Based Optimisation (Minimisation)

For a point x^* to be a local minimum point of a function $f(x) = f(x_1, x_2, ..., x_n)$, the conditions are that the gradient of the function at this point should be zero and the Hessian of the function at this point should be positive definite.

$$\nabla_{\boldsymbol{x}} f = 0 \text{ at } x = x^*$$

$$\nabla_{\boldsymbol{xx}} f \rightarrow \text{positive definite at } x = x^*$$

4.1 Positive definiteness

A symmetric matrix is said to be positive definite if that matrix has all its eigenvalues strictly positive.

4.2 Example

Problem statement: Find the minimum of the function $f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 3)^2 + 1$.

Solution:

$$f(\boldsymbol{x}) = (x_1 - 2)^2 + (x_2 - 3)^2 + 1$$

Gradient of the function is

$$\nabla_{\boldsymbol{x}} f = \begin{Bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{Bmatrix} = \begin{Bmatrix} 2(x_1 - 2) \\ 2(x_2 - 3) \end{Bmatrix}$$

Setting the gradient to zero implies

$$\begin{Bmatrix} 2(x_1 - 2) \\ 2(x_2 - 3) \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$$

$$\Rightarrow \begin{Bmatrix} x_1 - 2 \\ x_2 - 3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$$

$$\Rightarrow x_1 = 2 \text{ and } x_2 = 3$$

.

Now, Hessian of the function is

$$\nabla_{\boldsymbol{xx}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

.

Since the Hessian matrix is a constant matrix, it would be the same at every point – including the point $(x_1, x_2) = (2, 3)$ – which is $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.

We can observe that it is a diagonal matrix. Hence, the eigenvalues of this matrix would be 2 and 2, both of which are positive. Thus, the matrix is a positive definite matrix.

Since the gradient at $(x_1, x_2) = (2, 3)$ is zero and the Hessian at $(x_1, x_2) = (2, 3)$ is a positive definite matrix, the point $(x_1, x_2) = (2, 3)$ is a local minimum of the function. Since this is the only existing minimum of the function, it is also the global minimum of the function.

5 Gradient Descent

The idea behind Gradient Descent is that, since the gradient provides the direction through which the function value increases (at its infinitesimal neighbourhood), a small step in the opposite direction of the gradient is expected to decrease the function value. This slight decrease in the function is iteratively carried out until the function practically stops decreasing further (i.e., when it reaches a point where the gradient is close to zero).

For a function $f(\boldsymbol{x})$, in order to find a local minimum,

1. Take an initial guess.

2. Compute the gradient of the function at that point.

3. Take a very small step (of a length of α) in the opposite direction (descent direction) of the gradient. The resulting point is expected to have a less function value than that at the initial guess.

4. Now, starting from this new point, keep on repeating the steps 2 and 3, until the gradient becomes close to zero.

5. When the gradient is close to zero, stop the algorithm. The resulting point is expected to be a local minimum (or close to a local minimum).

To summarise, an initial guess $\boldsymbol{x}^{(0)}$ is chosen first, and then iterations are carried out as per the formula shown below:

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(n)})$$

The iterations are performed until a termination criterion is reached. One possible termination criterion could be that the norm of the gradient at the final point should be less than a specified limit.
This method of iteratively finding the local minimum starting from an initial guess, is known as Gradient Descent algorithm. The convergence of this method depends on what amount of step (α) is used. This parameter, i.e., α, is often referred to as the *learning rate*. For convergence, a smaller learning rate is recommended, however, smaller learning rates could require very large number of iterations, thereby significantly slowing down the process of finding of the local minimum.

5.1 Example

Problem statement: Suppose $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$. Starting from the point $\boldsymbol{x}^{(0)} = \begin{Bmatrix} 1 \\ 1 \end{Bmatrix}$, implement Gradient Descent algorithm with a learning rate of $\alpha = 0.01$ to find a local minimum of the function.

Solution:

$$f(\boldsymbol{x}) = f(x_1, x_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$$

$$\Rightarrow \nabla_{\boldsymbol{x}} f = \begin{Bmatrix} 2\left(x_1^2 + x_2 - 11\right) \cdot 2x_1 + 2\left(x_1 + x_2^2 - 7\right) \\ 2\left(x_1^2 + x_2 - 11\right) + 2\left(x_1 + x_2^2 - 7\right) \cdot 2x_2 \end{Bmatrix}$$

$$\boldsymbol{x}^{(0)} = \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \end{Bmatrix}$$

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha \nabla_{\boldsymbol{x}} f(x^{(0)}) = \begin{Bmatrix} 1 \\ 1 \end{Bmatrix} - 0.01 \begin{Bmatrix} -46.0 \\ -38.0 \end{Bmatrix} = \begin{Bmatrix} 1.46 \\ 1.38 \end{Bmatrix}$$

$$\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \alpha \nabla_{\boldsymbol{x}} f(x^{(1)}) = \begin{Bmatrix} 1.46 \\ 1.38 \end{Bmatrix} - 0.01 \begin{Bmatrix} -51.003 \\ -35.045 \end{Bmatrix} = \begin{Bmatrix} 1.97 \\ 1.73 \end{Bmatrix}$$

$$\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} - \alpha \nabla_{\boldsymbol{x}} f(x^{(2)}) = \begin{Bmatrix} 1.97 \\ 1.73 \end{Bmatrix} - 0.01 \begin{Bmatrix} -46.533 \\ -24.866 \end{Bmatrix} = \begin{Bmatrix} 2.435 \\ 1.979 \end{Bmatrix}$$

$$\vdots$$

The points and their derivative values over the iterations are shown in Table 1, and the plots for the two sets of values are shown in Figure 1 and Figure 2, respectively. From the plots, it can be observed that as the iterations are performed, the initial guess of $(x_1, x_2) = (1, 1)$ is approaching the point $(x_1, x_2) = (3, 2)$, which is a local minimum to this function. It can also be observed from Table 1 that near the point $(x_1, x_2) = (3, 2)$ the gradient is very close to zero.

Figure 1: The values of x_1 and x_2 over the iterations

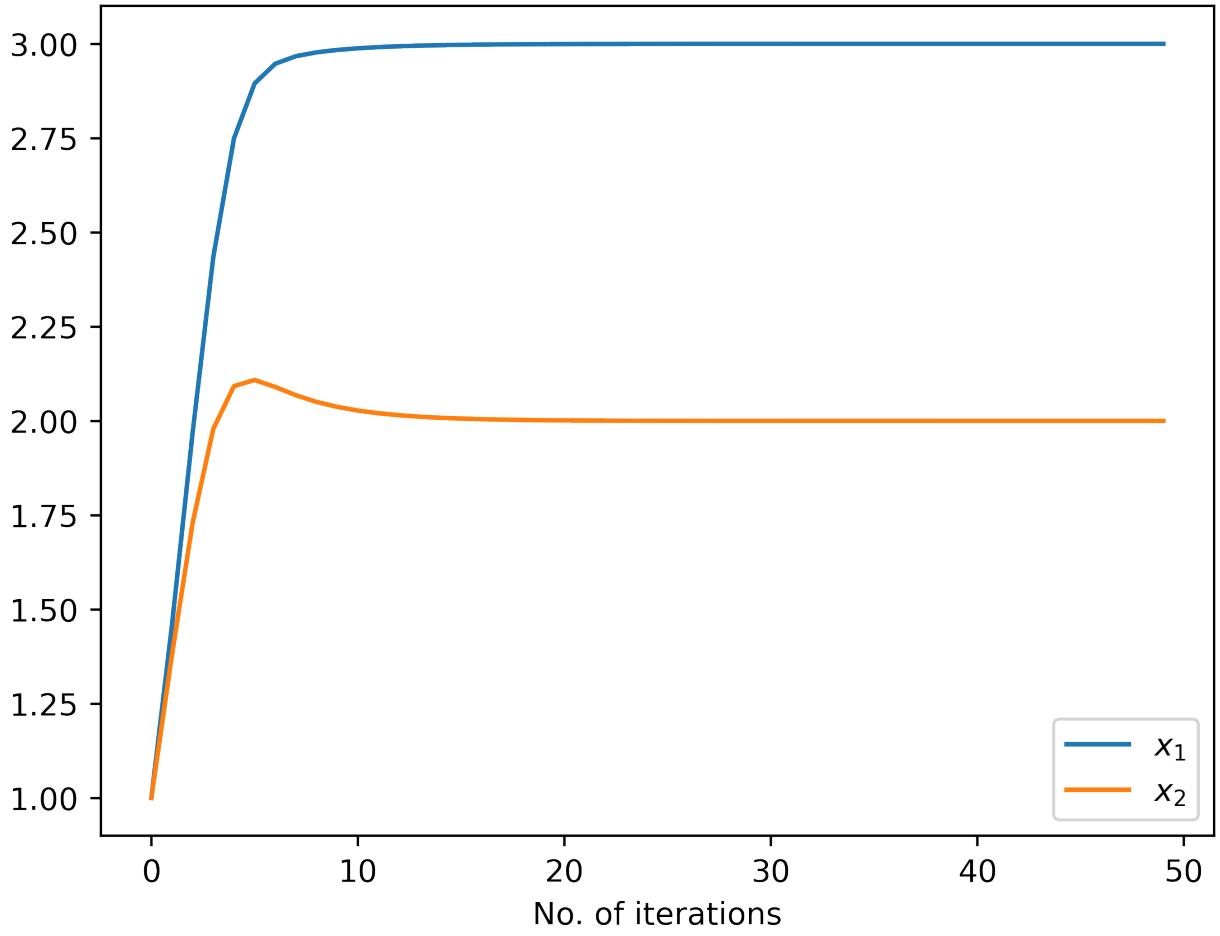

Table 1: Values at each iteration

S.No.	x_1	x_2	$\frac{\partial f}{\partial x}$	$\frac{\partial f}{\partial y}$
1	1	1	-46.0	-38.0
2	1.46	1.38	-51.00345600000001	-35.045312
3	1.97003456	1.73045312	-46.533204200441595	-24.8663530044276
4	2.435366602004416	1.979116650044276	-31.395354074751864	-11.307484049535772
5	2.7493201427519343	2.092191490539634	-14.582680493394037	-1.6387311396143431
6	2.8951469476858747	2.1085788019357774	-5.218331628419662	1.8591321157103127
7	2.947330263970071	2.0899874807786745	-2.0012906300065154	2.190030064655201
8	2.9673431702701363	2.0680871801321223	-1.016230557880558	1.7675897311734392
9	2.977505475848942	2.050411282820388	-0.6376518462130498	1.3220728075393895
10	2.9838819943110724	2.037190554744994	-0.4392171069758284	0.9736416327511448
11	2.9882741653808305	2.027454138417483	-0.31346584665638844	0.7160835986490245
12	2.991408823847394	2.020293302430993	-0.22710016109350661	0.527516066437647
13	2.993679825458329	2.015018141766616	-0.16582166469911286	0.3893401683479487
14	2.99533804210532	2.0111247400831367	-0.12166800177756443	0.2878137637467601
15	2.9965547221230957	2.008246602445669	-0.08956899417519892	0.21302341628030827
16	2.9974504120648477	2.006116368282866	-0.06609575043947657	0.15781414115226755
17	2.9981113695692425	2.004538226871343	-0.04885882522625806	0.11699462067036706
18	2.9985999578215052	2.0033682806646396	-0.03616312695188029	0.08677819673213304
19	2.998961589091024	2.002500498697318	-0.02679150028998123	0.06439063033028362
20	2.9992295040939236	2.0018565923940153	-0.019862307213841746	0.04779244066135506
21	2.999428127166062	2.0013786679874017	-0.01473280956433598	0.035480383170849894
22	2.9995754552617053	2.001023864155693	-0.010932181358509041	0.026344271740270517
23	2.9996847770752906	2.0007604214382906	-0.008114292957813518	0.019562989866753214
24	2.9997659200048687	2.000564791539623	-0.006024007181199664	0.014528549674085851
25	2.999826160076681	2.000419506042882	-0.004472885294447337	0.010790399667647728
26	2.9998708889296255	2.0003116020462057	-0.003321544917568530 8	0.008014450998813736
27	2.999904104378801	2.0002314575362177	-0.0024667758308503096	0.005952859208090139
28	2.9999287721371095	2.0001719289441366	-0.0018320901953039481	0.004421697454941028
29	2.9999470930390624	2.0001277119695873	-0.0013607693556814127	0.003284437774759382
30	2.9999607007326192	2.0000948675918395	-0.0010107352632537925	0.002439716950943183
31	2.999970808085252	2.00007047042233	-0.0007507608633297025	0.001812268727247985
32	2.9999783156938853	2.0000523477350574	-0.0005576660837789404	0.0013461990370038028
33	2.9999838923547233	2.0000388857446874	-0.0004142409975500527	0.0009999967179853686
34	2.9999880347646988	2.0000288857775077	-0.0003077064218537373	0.0007428306584865647
35	2.9999911118289173	2.0000214574709227	-0.00022857223970262908	0.000551801035039744
36	2.9999933975513144	2.0000159394605723	-0.00016979033479103123	0.00040989844958104585
37	2.999995095454662	2.0000118404760765	-0.000126125919393871 3	0.00030448846040008346
38	2.999996356713856	2.0000087955914725	-9.369084081001233e-05	0.0002261861422518873 7
39	2.9999972936222643	2.00000653373005	-6.959707311082752e-05	0.000168020235454809 3
40	2.9999979895929956	2.0000048535276953	-5.16994108394897e-05	0.0001248123359693357 4
41	2.999998506587104	2.0000036054043355	-3.8404382841861207e-05	9.271578439246255e-05
42	2.9999988906309323	2.0000026782464917	-2.8528334404143152e-05	6.887316209215273e-05
43	2.9999991759142763	2.0000019895148706	-2.1192020319915628e-05	5.116188092494275e-05
44	2.9999993878344795	2.0000014778960615	-1.5742313053124235e-05	3.800520523370394e-05
45	2.99999954525761	2.000001097844009	-1.1694048819066438e-05	2.8231875854617485e-05
46	2.9999996621980984	2.0000008155252504	-8.686831368009962e-06	2.0971835570696377e-05
47	2.9999997490664123	2.000000605806895	-6.452945207513494e-06	1.557877099473626e-05
48	2.9999998135958643	2.000000450019185	-4.793521007684376e-06	1.1572574173116067e-05
49	2.9999998615310743	2.000000334293443	-3.560830911908397e-06	8.59660109291738e-06
50	2.9999998971393835	2.0000002483274324	-2.6451365862101284e-06	6.385921769912944e-06

Figure 2: The values of $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_1}$ over the iterations